The " Weather Intelligence for Renewable Energies " Benchmarking Exercise on Short-Term Forecasting of Wind and Solar Power Generation

Simone Sperati, Stefano Alessandrini, Pierre Pinson, Georges Kariniotakis

Energies **2015**, *8*, 9594-9619; doi:10.3390/en8099594

OPEN ACCESS

energies

ISSN 1996-1073

www.mdpi.com/journal/energies

Article

The "Weather Intelligence for Renewable Energies" Benchmarking Exercise on Short-Term Forecasting of Wind and Solar Power Generation

Simone Sperati [1,*], Stefano Alessandrini [2], Pierre Pinson [3] and George Kariniotakis [4]

[1] Sustainability & Energy Sources, Ricerca Sistema Energetico (RSE) SpA, Milano 20134, Italy

[2] Research Applications Laboratory, National Center for Atmospheric Research (NCAR), Boulder, CO 80301, USA; E-Mail: alessand@ucar.edu

[3] Electrical Engineering, Technical University of Denmark (DTU), Kgs. Lyngby 2800, Denmark; E-Mail: ppin@elektro.dtu.dk

[4] Renewable Energies & SmartGrids, MINES ParisTech, Sophia Antipolis 06904, France; E-Mail: georges.kariniotakis@mines-paristech.fr

* Author to whom correspondence should be addressed; E-Mail: simone.sperati@rse-web.it; Tel.: +39-02-3992-5337.

Academic Editor: Guido Carpinelli

Received: 23 June 2015 / Accepted: 20 August 2015 / Published: 3 September 2015

Abstract: A benchmarking exercise was organized within the framework of the European Action Weather Intelligence for Renewable Energies ("WIRE") with the purpose of evaluating the performance of state of the art models for short-term renewable energy forecasting. The exercise consisted in forecasting the power output of two wind farms and two photovoltaic power plants, in order to compare the merits of forecasts based on different modeling approaches and input data. It was thus possible to obtain a better knowledge of the state of the art in both wind and solar power forecasting, with an overview and comparison of the principal and the novel approaches that are used today in the field, and to assess the evolution of forecast performance with respect to previous benchmarking exercises. The outcome of this exercise consisted then in proposing new challenges in the renewable power forecasting field and identifying the main areas for improving accuracy in the future.

Keywords: short-term energy forecasting; wind power; solar power; renewable energies; probabilistic forecasting; meteorological modeling; benchmarking comparison

1. Introduction

The renewable energy production share is constantly increasing due to both the scarcity of fossil fuel resources and to some strategic incentives made at global level to reduce carbon emissions in the atmosphere. Wind and solar energy play the main role in this process. The related energy production is highly variable in time as well as the weather processes (solar radiation and wind speed) to which they are dependent. Accurate and specific forecasting can help limiting the strong fluctuations, potentially induced in the electricity grid, by facilitating the balancing.

Concerning wind and solar energy forecasting, one can generally distinguish between deterministic and probabilistic approaches [1]. A deterministic approach consists in estimating a unique value of the variable to be predicted for each time-step in the future. A probabilistic approach focuses on informing about the full range of potential events in terms of power generation, for example through a set of conditional probability density functions (PDF) or a set alternative scenarios. In particular, probabilistic predictions can be based either on ensemble models, in which a model is run different times from slightly perturbed initial conditions [2], or otherwise produced using statistical methods (e.g., quantile regression [3]). This provides both a prediction about the probabilities of occurrence of a given event (*i.e.*, produced power greater than a given threshold) and information about the expected uncertainty affecting any issued single value forecast. While deterministic approaches for renewable energy forecasting have been explored for almost 30 years, probabilistic forecasting gained attention only in the last 10 years. By now it is becoming widespread, especially regarding wind energy.

The first publication regarding wind power forecasting appeared in 1984 [4]. Afterwards, in the following decades, an increasing amount of literature has been dedicated to the subject, following the rapid growth of power plant installations in many countries of the world. The most complete reviews of the state of the art in wind power forecasting can be consulted in [5,6], while a range of the most recent deterministic and probabilistic wind power forecasting applications is addressed in [7–9]. Regarding solar energy forecasting, the first attempts made for predicting solar irradiance can be traced back to [10]. Comprehensive reviews of the status of forecasting solar irradiance on different time scales for energy generation are reported in [11,12], while different forecasting techniques to predict solar power output are evaluated and compared in [13,14]. In addition to these references, an interesting review of a wide range of forecasting tools as statistical and computational intelligence models, focused on electricity price forecasting, can be found in [15]. Finally, in terms of comparison between model performance and forecasting approaches, [16–18] provide a useful look at what has been done regarding wind power forecasting.

The European Cooperation in Science and Technology (COST) Action ES1002 Weather Intelligence for Renewable Energies ("WIRE") gathers members from 27 European countries and five non-COST institutions in USA, Canada, Australia and Japan [19]. Its aim is to satisfy the requirements to provide the best possible specific weather information for forecasting the energy production of wind and solar power plants, especially in the short-term horizon (*i.e.*, from 0 h to 72 h ahead). Therefore, the team proposed to organize a benchmarking exercise aiming to validate the performance of a range of modeling approaches for renewable generation forecasting [20]. Nowadays, with the constant growth of renewable forecasting methods and technologies as well as operational tools, it becomes increasingly important to define a standardized set of procedures to be used when evaluating and

comparing results of different forecasting approaches applied to different energy sources. Beyond the standardized procedures proposed, the paper provides well-established reference performance values, where researchers and industrials may use as reference to situate the performance of new developments on forecasting models. This novel work was then organized aiming to propose a starting point towards the development of an international benchmarking platform, suitable for all possible applications in the renewable energy forecasting field. In fact, it is the first time that a wide range of both wind and solar, deterministic and probabilistic power forecasting techniques are evaluated and compared on different, real test cases with varying topographical and meteorological conditions. For the first time, several wind and solar prediction methods are evaluated under a common framework at different locations. This allows understanding potential differences in the predictability of the two resources, which is an information of primary importance for power system operators. On the other hand it is demonstrated in the paper that a common evaluation framework, when thoroughly defined, can be applicable for both wind and solar forecasting. The other novelty aspect of the paper is to evaluate the improvement achieved by wind power forecasting in the last 10 years. In fact, for one of the test cases, a comparison between the performance obtained with new modeling techniques and past results obtained in the frame of the ANEMOS EU project [18] is addressed. The exercise was performed on four different test cases for wind and solar power plants in Europe. The available historical data sets include measurements of power output and meteorological variables from the plant monitoring sites, as well as numerical weather prediction (NWP). The participants had access to historical data sets to train their prediction models, but the organizers, to evaluate the submitted forecasts, withheld parts of the data. The participants were free to use any modeling approach, as well as their own numerical weather forecasts (either generated by own models or by weather services). The goal was to produce 3-day ahead point forecasts, with hourly or three-hourly time-steps. Probabilistic forecasts were also requested in form of quantiles of the predicted distribution. The submission of entries included the forecasted values of the power output, as well as a short description of the modeling approach and of the additional inputs, if used.

The evaluation and analysis of the results is based on an updated version of a standardized protocol defined during ANEMOS [21]. Regarding probabilistic forecasts, some of their properties are evaluated on the basis of the framework proposed in [22]. The exercise was well received and gathered participants from all-over Europe and from Japan, India, Australia and USA.

Finally, it should be noticed that some of the results presented in this paper were presented at an early stage in [23]. In this work, the authors are presenting a more comprehensive evaluation covering every aspect of the exercise, adding also a whole new investigation of potential improvement achievable by using higher resolution forecasts on one of the test cases.

This paper is organized as follows: Section 2 provides a description of the exercise and the test cases; Section 3 describes the evaluation framework; Section 4 presents the results, which are then discussed in Section 5; and conclusions are drawn in Section 6.

2. Description of the Benchmarking Exercise: Setup and Data

Data for two wind farms (Abruzzo, Klim) and two PV plants (Milano, Catania) were used. The power plants were chosen in order to consider different meteorological and topographical conditions, as noticeable in the map of Figure 1 in which their location is shown.

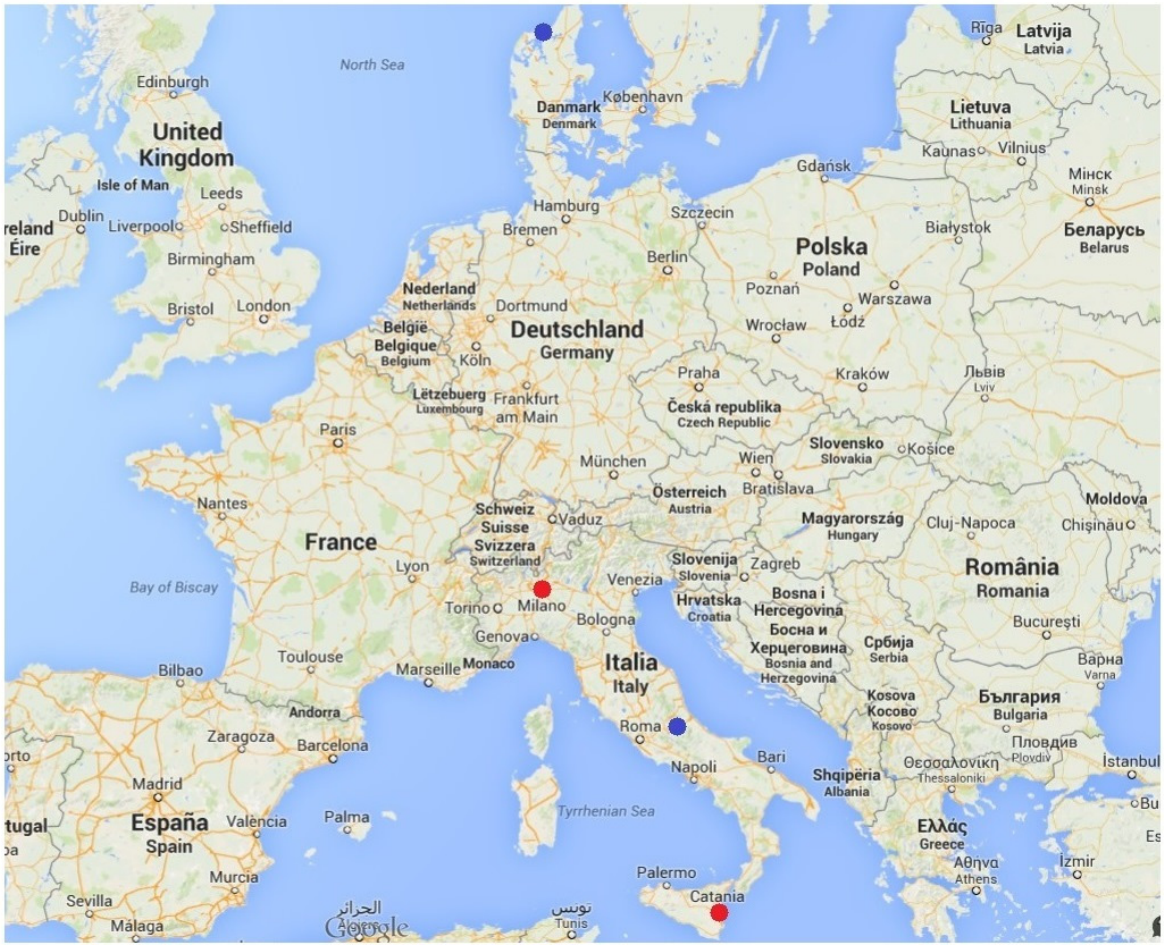

Figure 1. Locations of the wind farms (blue dots) and the PV plants (red dots).

The data sets cover a period of 1.5 or 2 years depending on the case. The organizers, using in each case different models or configurations, also provided meteorological data. Details are thoroughly explained in the following paragraphs.

2.1. General Design Aspects

Data used for the exercise are not public. For this reason, Non-Disclosure Agreements were signed by all the participants and the providers controlled the exchange of data and information. After that, data were provided all at once to the participants. The exercise was announced in January 2013, allowing participants to apply until March 2013 with three months to complete it. The deadline for submission of results was set at the end of June 2013. The organizers evaluated the results after collecting all the submissions. Overall, 33 participants applied for the exercise. They are mainly from research institutions, meteorological services or universities all-over Europe, with also some from Japan, India, Australia and USA. Also, a few participants are from commercial companies. Only 18 deliveries of predictions out of 33 initial applications were received. Most of the participants applied only for the wind or for the PV part of the exercise, only a few for both parts. Just three participants delivered probabilistic forecasts for the wind power part while nobody sent solar probabilistic forecasts.

In order to allow the participants to train their models and to have a common validation period, the data sets were split into a training and a test period as shown in Table 1.

Table 1. Training/test period definitions for the different data sets.

Power plant	Training period	Test period
Abruzzo	1 January–31 December 2010	1 January–31 December 2011
Klim	1 January–31 December 2001	1 January–31 December 2002
Milano	1 July–31 December 2010	1 January–31 December 2011
Catania	1 January–31 December 2010	1 January–31 December 2011

Except for Milano, where a 6-month training period was defined, the other cases had a 1-year period available for training purposes. All power and forecast data were available to the participants for the training periods, while for the test periods measured power data were masked for the first 14 days of each month. The 14 day masking was imposed in order to reduce potential cheating by the participants. No observations were made available to the participants during the masked periods that could eventually be used to adjust and unfairly improve predictions. At the same time, the authors wanted to allow a rolling month by month retraining of the prediction models, keeping a static period where observations were not provided. These masked periods were used by the organizers for the evaluation, discarding forecasts received for the remaining days of each month.

2.2. Wind Power

Abruzzo and Klim wind farms were used as test cases for the wind power prediction part. As previously stated, they were chosen due to the great difference in their meteorological and topographical conditions.

The Abruzzo wind farm is located in a complex-terrain area in a central region of Italy, and has a nominal power (*NP*) around 100 MW (for confidentiality agreement, the authors are not allowed to disclose exact information about the wind farm). Hourly power data for the period 01 January 2010–31 December 2011 were provided to the participants. Meteorological data from the European Centre for Medium-Range Weather Forecasts (ECMWF) deterministic model were also provided, using 0–72 h ahead runs starting at 12 UTC with 3-hourly frequency and 0.125° horizontal resolution. The forecasts were computed interpolating the gridded data at the wind farm position. The meteorological variables provided are wind speed and wind direction at 10 m height, temperature at 2 m height, boundary layer height and mean sea level pressure.

The Klim test case considers an on-shore wind farm in a flat terrain area of Northern Denmark, with a *NP* equal to 21 MW. Hourly power data were provided for the period 1 January 2001–31 December 2002. This data set was already used for comparison in [18]. The same meteorological data by the Danish HIRLAM model used in [18], with 0.15° spatial resolution and hourly time-steps, were provided to the participants. Data were interpolated at the wind farm position. HIRLAM was nested on the ECMWF meteorological fields, which serve as initial and boundary conditions, performing a 0–48 h ahead run every 6 h (*i.e.*, at 0, 6, 12 and 18 UTC). Wind speed, wind direction and temperature were extracted at model levels 30 and 31 (*i.e.*, corresponding to about 148 m and 30 m respectively). The ECMWF deterministic model configuration used at that time was TL511, with about 40 km spatial resolution.

The use of lead time 0 in the wind forecasts represents the power predictions obtainable by using the meteorological analysis and can be useful to define a predictability limit at any wind farm. In fact, the meteorological analysis can be seen as the best meteorological forecasts available, not being

affected by errors depending on the lead time. Besides the different complexity of terrain between the two test cases, it should be also noticed the higher availability of wind resource in Denmark rather than in Italy.

2.3. Solar Power

Similarly for wind power, the solar power prediction part of the exercise involved two PV plants representative of different meteorological conditions.

The first PV plant has a *NP* of 5.21 kW. The plant is located in the suburbs of the city of Milano, in Northern Italy. Hourly power data for the period 1 July 2010–31 December 2011 were available to the participants, along with measured global horizontal irradiance (GHI), direct normal irradiance (DNI) and temperature at 2 m height. Forecasts data of GHI, total cloud cover (TCC) and temperature at 2 m height were obtained by the ECMWF deterministic model by a bilinear interpolation at the plant location. The forecast runs were initialized at 12 UTC, with 3 h time-steps for the period 0–72 h ahead, and a grid spacing equal to 0.125°.

The other PV plant is located in the suburban area of Catania, in Southern Italy, where solar irradiance availability is generally higher than in Northern Italy. The plant is made by the same photovoltaic components as the Milano plant. Hourly measured power, GHI, DNI and 2 m temperature data for the period 1 January 2010–31 December 2011 were available. Forecasts performed with the Regional Atmospheric Modeling System (RAMS) [24] were also provided. RAMS was configured with two nested grids, with a horizontal resolution of 15 km and of 5 km, producing forecasts of GHI, DNI, TCC and 2-m temperature. In the model configuration, the Harrington parameterization [25] was used as the radiation scheme. Bulk microphysics parameterization was also activated in order to account for full moisture complexity. The model was run starting at 0 UTC with hourly frequency and nested on 6 h ECMWF boundary conditions with 0.125° spatial resolution.

3. Evaluation Framework

A common verification framework based on [21] is adopted in order to evaluate the forecasts. The mean absolute error (*MAE*) is used as a ranking criterion. The index can be expressed as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|o_i - f_i| \quad (1)$$

where o_i is the ith observed value and f_i the ith forecasted value. *MAE* allows measuring the average error magnitude in the forecasts.

Another common verification index is the root mean squared error (*RMSE*). However, being a quadratic score index, RMSE often gives relatively higher weights to larger forecast errors. *MAE*, on the other hand, is a linear score, *i.e.*, all the individual errors are given the same weight in the average. *MAE* values are normalized by *NP* and reported as a percentage. The index is also normalized by the mean power (*MP*) data measured during the test period. In fact, *MAE* is usually proportional to *MP*. The normalization by *MP* allows then a fairer comparison of the model performance at sites with different terrain, meteorological conditions and wind resource availability.

Model bias is also calculated over the whole forecast horizon and reported. Bias corresponds to the systematic error and is expressed by the following formula:

$$bias = \frac{1}{N}\sum_{i=1}^{N}(o_i - f_i) \tag{2}$$

In order to perform an additional evaluation of the best result selected with the *MAE* criterion, and also to check for statistically significant differences with the second best, the Diebold-Mariano (DM) test [26] is computed. The DM test compares the forecast accuracy of two forecasting methods testing the null hypothesis H0 (*i.e.*, the two competing forecasts have the same level of accuracy) *versus* the alternative hypothesis H1 (*i.e.*, the second result is actually less accurate than the first one). With a *p*-value resulting from the test close or equal to 1 it is difficult to draw conclusions and one simply cannot reject the null hypothesis, while very low *p*-values allow accepting H1. DM test is applied on the residuals putting the data over all the lead times together.

Error distributions are also analyzed for the best forecasts of each test case.

The continuous ranked probability score (*CRPS*) is chosen for ranking the probabilistic forecasts. The *CRPS* is a common verification index that compares a full probabilistic distribution with the observations, when both are represented as cumulative distribution functions (CDF) [27], and it is expressed as follows:

$$CRPS = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\left({F_i}^f(x) - {F_i}^0(x)\right)^2 \mathrm{d}x \tag{3}$$

where ${F_i}^f(x)$ is the CDF of the probabilistic forecasts for the *i*th value at each time step and ${F_i}^0(x)$ is the CDF of the measurements. *CRPS* consists then in the mean squared error of the cumulative distribution, and can be reduced to the *MAE* for a deterministic forecast. The index is computed through numerical integration techniques, based on a discretization of the forecast CDF using its various defining quantiles, as well as the corresponding power measurement. A lower value of the *CRPS* means better performance. The *CRPS* has the same dimension as the forecasted variable.

Further verifications of probabilistic performance is made using rank histograms, which assess statistical consistency of a probabilistic distribution (*i.e.*, if the members of a probabilistic distribution are statistically indistinguishable from the observations). In other words, it is verified whether the measured values ranked among the corresponding ordered members equally take any rank in the whole range of the PDF. A perfect result is given by a flat rank histogram, in which the measurements have a uniform probability rank equal to $\frac{1}{n+1}$. Considering that there can be significant amount of cases with zero production associated with several members equal to zero, the ranks are assigned properly in the evaluation in correspondence of these cases. For each case in which that happens, ranking frequencies of the first *n* ranks (where *n* is the number of sorted members with zero value associated to a measurement equal to zero) are computed as $\frac{1}{n}$.

Sharpness, which is an important attribute of probabilistic forecasts, is also assessed with proper diagrams. Sharpness refers to the ability of a probabilistic system to issue forecasts towards the extreme probability classes, for any given threshold. In other words, plotting the relative forecast

frequency (*i.e.*, the number of cases) in each class interval of probability allows checking if the system is sharp.

Other indices can be of interest when evaluating probabilistic forecasts. In this context, the evaluation process is kept on the same metric asked to the participants during the exercise, when *CRPS* was proposed as main probabilistic ranking index. Having also results based on rank histograms and sharpness diagrams is then enough to assess the general performance of the different forecasts.

Finally, considering that the forecasts are available for three days ahead and, in the case of Klim, there are four runs per day, the authors made some verification checks in the submitted forecasts in order to avoid potential use of additional information by the participants, *i.e.*, future information not available at the moment forecasts are made, giving an advantage. In fact, one participant was excluded after checking that his forecasts for the 24–48 and the 48–72 h ahead horizons consisted in each case in the more recent forecasts for the 0–24 h ahead horizon, issued the following days. This was evident by a clearly recurring pattern in his error trends.

4. Evaluation Results

An id number is assigned to identify each participant in the evaluation, without any reference to their names. Since some participants provided forecasts for both the wind and solar parts of the exercise, the correspondences between ids referring to the same participant in both parts are identified in Table 2.

Table 2. List of participants with correspondences between wind and PV test cases.

Test case	**id correspondence**			
Wind	id01	id02	id04	id09
PV	id02	id03	id04	id06

All the forecasting methods used by the participants for both wind and solar test cases are summarized in Table 3 and further explained in this section.

Table 3. List of the forecasting methods used by the participants for each test case.

id	Wind		PV		Probabilistic
	Abruzzo	**Klim**	**Milano**	**Catania**	
01	Own meteorological model + Kalman Filter, Artificial Neural Network (ANN) and Ensemble Learning		WIRE model data + linear regression (random forest)		Own meteorological model + Kalman filter, ANN And ensemble learning
02	Combination of HIRLAM and GFS + wake parameterization based on atmospheric stability + power curve (density corrected) + ANN	WIRE data + ANN	Own meteorological model + output correction using tendency of past production		-
03	WIRE data + non-linear function approximation between wind speed and direction to wind power output		GFS + Model Output Statistics + conversion to power	WIRE data + conversion to power	-
04	WIRE data + Support Vector Machines		WIRE data + Support Vector Machines		-
05	WIRE data + ANN		Linear regression between GHI and solar power		-
06	WIRE data + feed-forward multilayer perceptron ANN		WIRE data + ANN (Multilayer Perceptron with Standard Back Propagation and Logistic Functions)		-
07	WIRE data + power curve obtained by linear interpolation between fitting power values		WIRE data + quantile regression to estimate clear sky production, irradiation and medium temperature + linear regression to explain the rate of clear sky production observed		-
08	Computational Fluid Dynamics (CFD) model		WIRE data + linear regression model		-
09	WIRE data + ANN (Multilayer Perceptron with Standard Back Propagation and Logistic Functions)		-	Combination of WIRE data and WRF ARW model version 2.2.1 using initial and boundary conditions from NCEP GFS + Gaussian Generalized Linear Model	-
10	WIRE data + average of ensemble ANN initialized with different weights		-		WIRE data + conditional kernel density estimation with a quantile-copula estimator
11	WIRE data + combination of time series and approximation		-		-
12	Hybrid approach combining physical modeling and advanced statistical post-processing (including a combination model applied on different prediction feeds)		-		-
13	-		-		WIRE data + local quantile regression using wind speed and wind direction as predictors

4.1. Deterministic Wind Power Forecasts

Figure 2 shows the *MAE* trends for the Abruzzo test case, whose evaluation is carried out on 0–72 h ahead forecast lead times with 3-h time-steps. *MAE* values are also calculated on the whole forecast range for the ranking, and are reported in Figure 2.

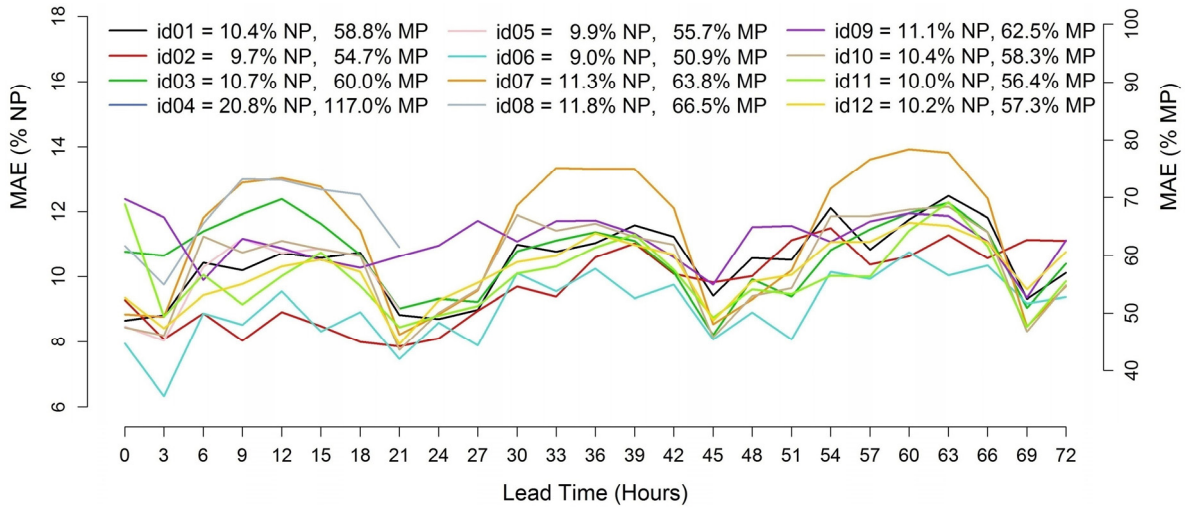

Figure 2. Error trends of the 0–72 h ahead wind power forecasts starting at 12 UTC for the Abruzzo test case (*MAE*/*NP* % and *MAE*/*MP* %). *MAE*: mean absolute error; *NP*: nominal power; and *MP*: mean power.

Except for a clear outlier, similar *MAE* trends are observed, with slightly higher errors for larger lead times. There is, however, a significant difference between the lowest and the highest *MAE*, which span in about 4 or 5 *MAE*/*NP* percentage points, depending on the lead time. Almost all the models show a strong daily cycle with larger *MAE* values during evening and night hours. Looking at the *MAE*/MP values, it can be noticed that most of the forecasts range between 50% and 65%, depending on the forecast horizon. The best result is achieved by id06 with a total *MAE*/*NP* of 9.0% and a *MAE*/*MP* of 50.9%. The DM test applied comparing id06 to the second best (id02, 9.7% *MAE*/*NP* and 54.7% *MAE*/*MP*) returns a *p*-value of 3.91×10^{-6}, allowing rejection of the hypothesis that the forecasts have the same accuracy.

The results for the Klim wind farm are shown in Figure 3. In this case, 0–48 h ahead, hourly forecasts are evaluated, considering all four initialization times (*i.e.*, 00, 06, 12 and 18 UTC) together.

The results show a best score of 9.5% *MAE*/*NP* and 43.7% *MAE*/*MP*, achieved again by id06. The outcome of 8.8% *MAE*/*NP* achieved by id04 is not considered, since the participant just provided forecasts for the 0–23 h ahead horizon without considering longer lead times. It is possible to observe a more defined trend with increasing *MAE* values for larger lead times than for Abruzzo, but with considerably lower dispersion of the error values (*i.e.*, about 1–2 *MAE*/*NP* percentage points, depending on the lead time). There are also a couple of outliers probably caused by some basic mistakes in the forecasting method. The DM test applied to id06 *versus* the second best (*i.e.*, id12, 9.6% *MAE*/*NP* and 44.2% *MAE*/*MP*) returns a *p*-value of 1.27×10^{-141}, meaning that even in this case the hypothesis of same forecast accuracy can be rejected. Overall, some considerations can be drawn with respect to the differences in the error trends between the two sites.

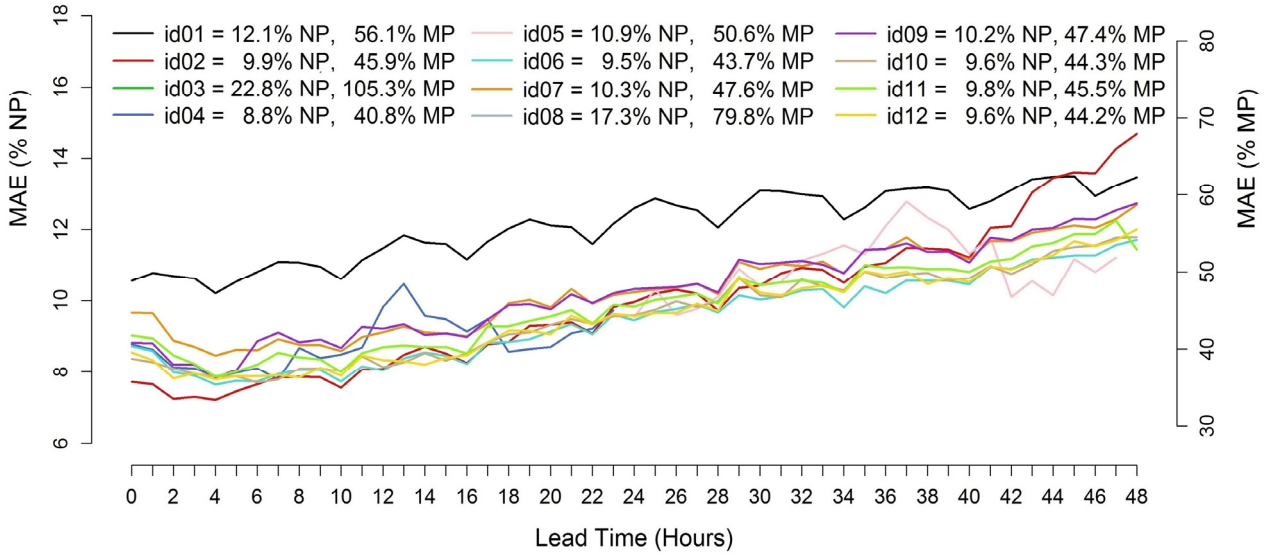

Figure 3. Error trends of the 0–48 h ahead wind power forecasts starting at 0, 6, 12 and 18 UTC for the Klim test case (*MAE*/*NP* % and *MAE*/*MP* %).

Abruzzo is characterized by a strong daily cycle of power production, which is not observed at Klim. Higher errors during night at Abruzzo are due to higher production during those hours, in fact the average wind power at 0 UTC is 6% higher than at 12 UTC. The authors believe that the increasing trend over lead time observed at Klim is related to the kind of error characterizing the meteorological forecasts. In fact, Klim is located on a flat terrain, where the meteorological forecasts are affected by lower representativeness errors (bad representations of topography, land use, kinematic winds *etc.*), which are higher for complex-terrain sites like Abruzzo. These kinds of error mask those caused by the decreasing predictability of the atmospheric flows over lead time more at Abruzzo than at Klim. Table 4 reports model bias with respect to *NP* for both Abruzzo and Klim. Bias is computed as average over the whole forecast period.

Table 4. Model bias for Abruzzo and Klim. Bias is expressed as percentage of NP.

Power plant	id01	id02	id03	id04	id05	id06	id07	id08	id09	id10	id11	id12
Abruzzo	1.6	0.9	2.4	2.1	1.1	0.7	5.3	0.9	0.4	0.7	0.5	4.4
Klim	−2.9	−0.5	0.0	3.2	0.0	−0.5	−2.5	2.3	−2.4	0.3	−1.5	1.9

It is possible to extract more information from the error distributions calculated from the best forecast of each test case. The diagrams, produced for the whole forecast range with bins representing 5% of *NP*, are reported in Figure 4.

The grey shaded area in each histogram delimits the 5%–95% quantiles interval. In both cases, a tendency to slightly overestimate wind power is noticeable from the positive skew in lower bins, especially for the Klim case. The distribution obtained for Abruzzo is slightly sharper than the one for Klim and shows prediction errors lower than 7.5% of *NP* 67% of the times, while for Klim this happens in 62% of the cases.

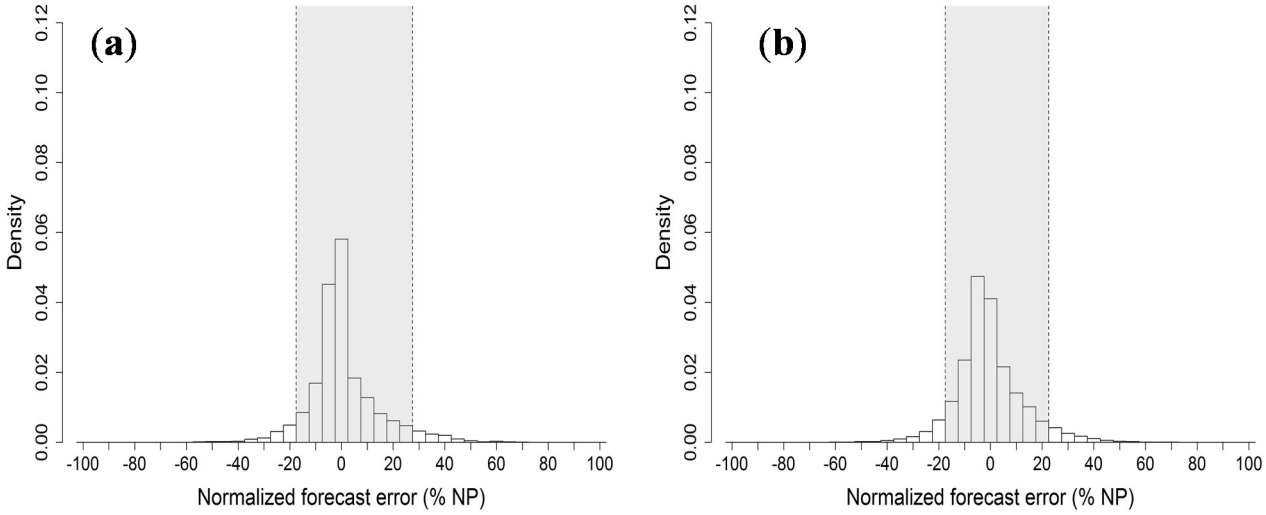

Figure 4. Error distributions (id06) of the best forecasts for the Abruzzo case (**a**) and the Klim case (**b**). The grey area in each histogram delimits the 5%–95% quantiles interval.

In both test cases id06 achieves the best result. His forecasting method is based on the use of meteorological data provided by COST and an application of two statistical approaches combining artificial neural networks (ANN) and generalized linear models (GLM) [28]. The two methods were used separately to learn the non-linear relation between the historical weather forecasts and the wind farm power measurements. The ANN consists in a feed-forward multilayer perceptron with one hidden layer and was optimized with the Levenberg-Marquardt algorithm. In the GLM, a logit function was applied as a link function and the response variable (*i.e.*, wind power) was assumed to be binomially distributed. Wind power production data was normalized between 0 and 1 by division with *NP*. The forecasts of all meteorological parameters were standardized and used as model input together with information about the time of day. ANN and GLM used the same set of input data. Before training the models, power data was filtered manually with respect to non-plausible values. The outputs of the ANN and GLM were then averaged to get a final wind power forecast for each wind farm.

Concerning the forecasting methods used by other participants, it should be noticed that most of them consist of different post-processing techniques applied to the NWP data provided by the organizers.

ANNs were used (e.g., id02, id05) with variable performance depending on the case. In the case of id10, forecasting the average of an ensemble of ANN initialized with different weights proves to be quite effective, ranking third on the Klim case. A method based on a combination of time series and approximation (id11) also led to good performance on both power plants.

Further methods were based on other machine-learning techniques, e.g., support vector machines (SVM): id04 applied SVM with fairly good results on Klim, providing forecasts only for the 0–24 h ahead interval; id01 chose to run a non-hydrostatic multi-nested model followed by a combination of Kalman Filter, ANN and Ensemble Learning techniques as post-processing, but with less effective results especially in the case of Klim; id07 used a fitting methodology that divided sample data into bins and computed a power series by minimizing the variance between power values inside the bins and the fitting power, obtaining a power curve by linear interpolation between the fitting power values. The approach was, however, not so effective, especially on Abruzzo.

A Computational Fluid Dynamics (CFD) model was applied by id08, but the lack of data regarding the wind farms (*i.e.*, single turbine data, anemometer measurements on site) didn't allow setting up a complete post-process and thus obtaining good results.

For Abruzzo, the second best result is obtained by id02 using additional meteorological data from the GFS global model, followed by application of a wind farm model (accounting for wake parameterization based on atmospheric stability) and reconstruction of a density corrected power curve. Data were also processed by an ANN. For Klim, the second best result is achieved by id12 with a hybrid approach combining physical modeling and advanced statistical post-processing, including a combination model applied on different prediction feeds.

4.2. Klim Case—Comparison with Previous Benchmark and High Resolution Model Run

As previously explained, forecast data for Klim are the same used back in 2002 and tested in [18]. HIRLAM spatial resolution at that time was equal to 0.15°, and it was driven by the ECMWF boundary conditions fields with 0.5° resolution. In Figure 5, a comparison between the new results obtained during the current exercise and the results obtained in [18] is addressed. It should be noticed that the amount of available data in [18] was higher. In particular, 2 years of data were used as a training period (from January 1999 to February 2001), while forecast evaluation was performed on data from March 2001 to April 2003.

Figure 5 compares the *MAE* as a function of forecast lead time of the best two forecasts from the previous benchmark and the current exercise. In this comparison data for lead time 0 is missing, in fact it wasn't evaluated at that time. As in Figure 3, the forecasts are compared considering all four initialization times together.

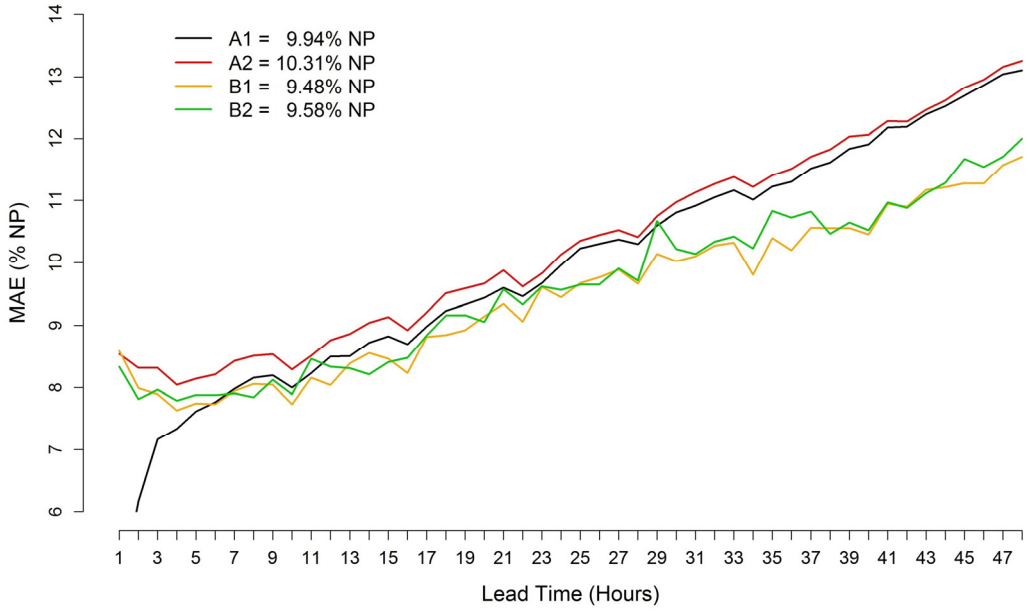

Figure 5. Comparison between new and old results in term of *MAE*/*NP* as a function of forecast lead time (1–48 h ahead) for the wind power forecasts issued at 0, 6, 12 and 18 UTC for the Klim test case. A1, A2 and B1, B2 are the best 2 forecasts of the previous benchmark and the current work respectively.

In the diagram, A1 and A2 are the best 2 forecasts issued for [18], while B1 and B2 correspond to the best forecasts of the current exercise. B1 and B2 are generally better in terms of *MAE*, especially for the forecast horizon from 24 h to 48 h ahead. During the previous benchmark some participants used models with auto-adaptive capabilities, which can benefit e.g., from using available online data. This appears evident looking at the A1 case, outperforming the others up to 5 h ahead. For longer lead times, however, the performance of A1 degrades and remains lower than those of B1 and B2. This is likely due to an improvement in the statistical post-processing techniques adopted by B1 and B2.

Furthermore, in order to investigate the potential improvement achievable by using higher resolution forecasts for the same power plant, the organizers have performed a new forecast run using RAMS. These forecasts were not provided for the exercise: however, a comparison with its results is addressed here.

ECMWF reforecast data with 0.25° horizontal resolution were retrieved for the same 2-year period. RAMS runs starting at 12 UTC using 2 nested grids with horizontal resolution of 12 km and 4 km were performed. Therefore, higher resolutions in both boundary conditions and the limited-area model were used. A post-processing system based on an ANN [29,30] was applied to both RAMS and HIRLAM output. As a consequence, the post-processing model is similar to the one used by the winner in the exercise, even if the GLM part was not performed. For this test the same conditions imposed on the participants were maintained (*i.e.*, missing data for the first 14 days of each month of the test period).

Figure 6 shows the *MAE*/*NP* and *MAE*/*MP* as a function of forecast lead time, calculated only for the 12 UTC model runs. The results obtained by all participants are shown in grey while the red line refers to the model chain RAMS + ANN. Results obtained by the ANN post-processing applied to HIRLAM data (HIRLAM + ANN) are also reported in blue.

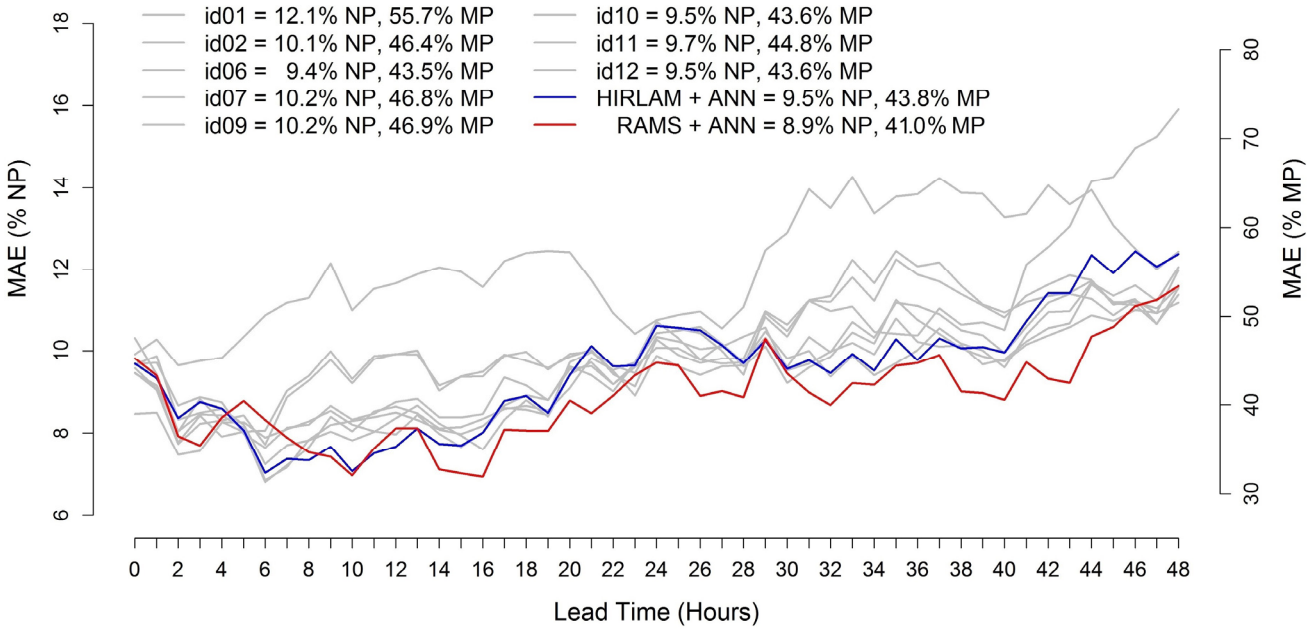

Figure 6. Error trends of the 0–48 h ahead wind power forecasts starting at 12 UTC for the Klim test case (*MAE*/*NP* % and *MAE*/*MP* %).

With 8.9% *MAE*/*NP* and 41.0% *MAE*/*MP*, RAMS + ANN allowed gaining about 0.5% of *MAE*/*NP* and 2.5% of *MAE*/*MP* on the best result observed during the exercise. The application of HIRLAM + ANN

allowed obtaining similar results as those obtained by other participants, with 9.5% *MAE*/*NP* and 43.8% *MAE*/*MP*. The application of the DM test on RAMS+ANN *versus* the result obtained by id06 returns a *p*-value of 9.2×10^{-2}, which allows rejecting the null hypothesis (*i.e.*, RAMS+ANN results to be better than id06).

The improvement shown by RAMS+ANN provides evidence of the benefits of using a higher spatial resolution both in the boundary conditions (which have been improved in the last 10 years due to the development carried out on the ECMWF deterministic model) and also in the limited area model. However, other model improvements like data assimilation schemes and physics parameterizations can also have contributed to increase the performance.

4.3. Deterministic Solar Power Forecasts

Solar statistics are computed on forecast data filtered using solar height (*i.e.*, forecasts in correspondence of lead times with solar height equal to zero are discarded). Figure 7 displays the *MAE* trends for Milano, calculated for the 3–72 h ahead forecast interval with 3-hourly time-steps. Two outliers behave very differently from the other models, especially during the 3–6 h ahead interval. Apart from these, a common trend between the different models is observed with the forecast errors reaching their peaks at 12 UTC of each forecast day. The difference between the error values of different models ranges between 2% and 3% of *MAE*/*NP*. The best score is obtained by id07 with an *MAE*/*NP* of 7.0% and an *MAE*/*MP* of 30.1%. The DM test applied comparing id07 with id01 (7.4% *MAE*/*NP* and 31.7% *MAE*/MP) returns a *p*-value of 0.961. It is thus difficult to state whether one model is actually more accurate than the other, and the null hypothesis cannot be rejected.

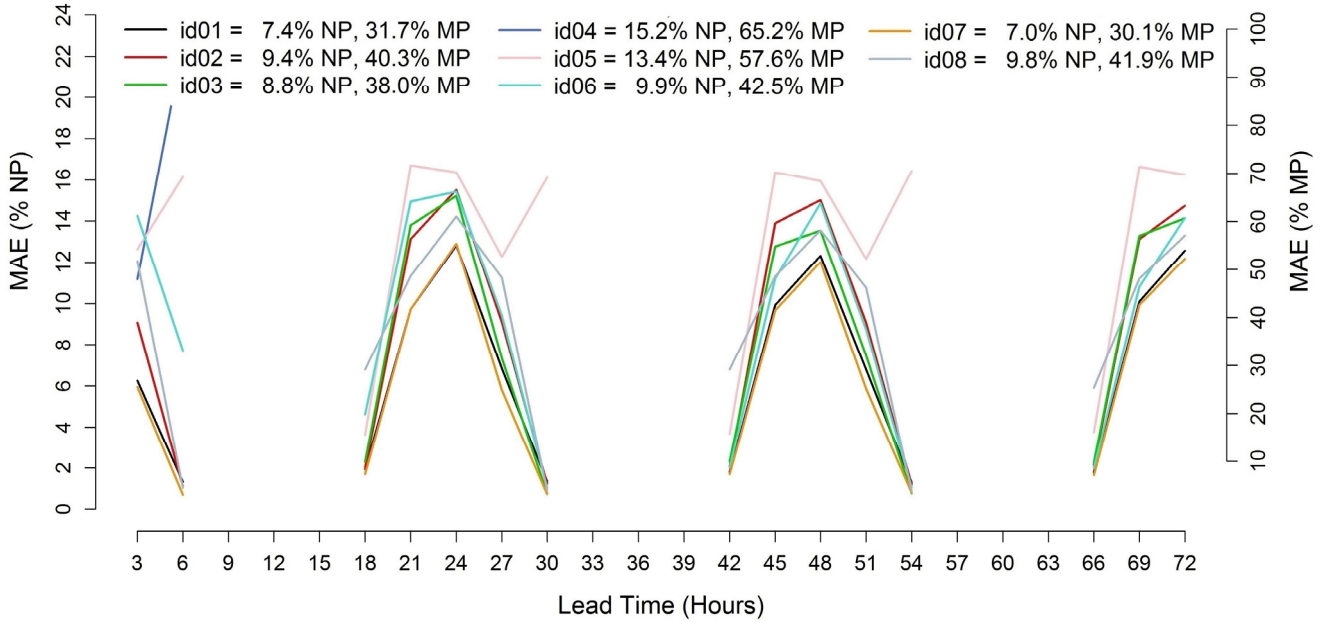

Figure 7. Error trends of the 3–72 h ahead solar power forecasts starting at 12 UTC for the Milano test case (*MAE*/*NP* % and *MAE*/MP %).

Figure 8 shows the *MAE* trend for the 1–72 h ahead, hourly forecasts made for Catania with hourly time-steps.

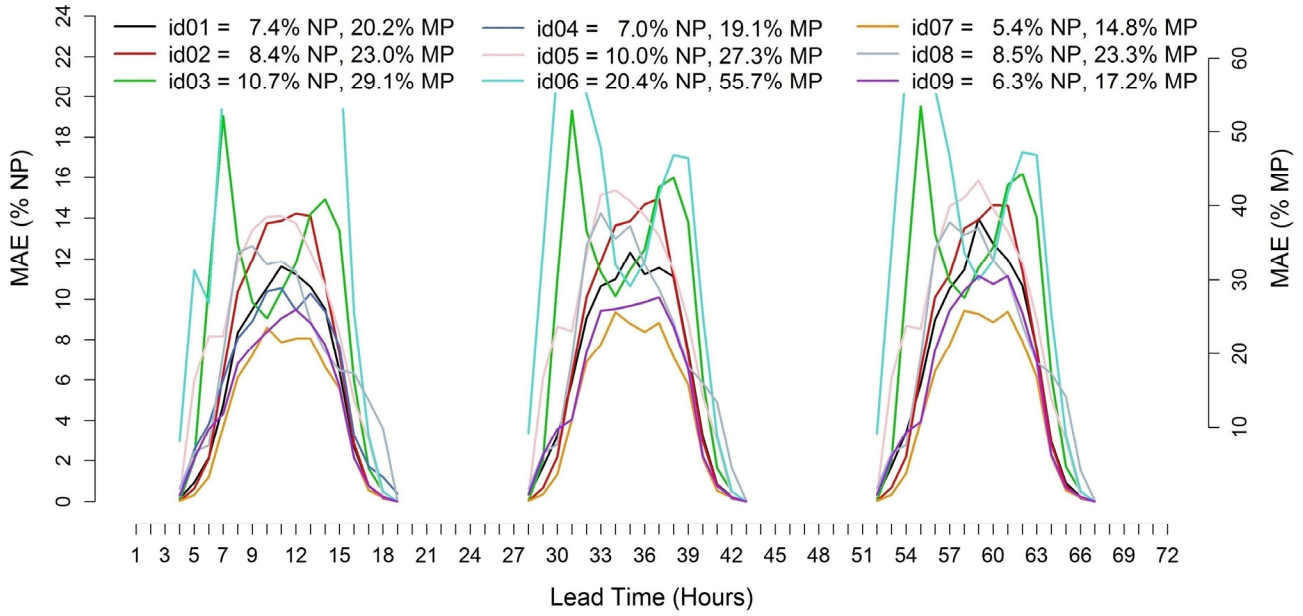

Figure 8. Error trends of the 1–72 h ahead solar power forecasts starting at 0 UTC for the Catania test case (*MAE*/*NP* % and *MAE*/*MP* %).

Except for a couple of models which exhibit two peaks during early morning and late afternoon hours, all the others show a maximum *MAE* around 12 UTC. The differences between the models are higher than for Milano test, reaching about 5%. id07 performed again better than the others with a 5.4% *MAE*/*NP* and 14.8% *MAE*/*MP*. The result is particularly good, with almost 1% *MAE*/*NP* less than the second best result of 6.3%, achieved by id09. However, the *p*-value of 0.999 returned from the DM test prevents rejecting the null hypothesis.

Looking at Figures 7 and 8 one could notice that solar power forecasting error trends are strongly dependent on the daily cycle. This is due to the solar elevation trend that partly masks lead time dependent errors. Also, the meteorological model's skill in forecasting solar irradiance and cloud coverage is not strongly dependent on lead time, and this is reflected on the power predictions. Bias values for both Milano and Catania are reported in Table 5.

Table 5. Model bias for Milano and Catania. Bias is expressed as percentage of *NP*.

Power plant	id01	id02	id03	id04	id05	id06	id07	id08	id09
Milano	−1.3	0.2	3.3	−10.6	0.9	−0.7	0.0	−2.2	-
Catania	1.5	3.8	0.6	−1.5	−1.7	8.7	−0.4	1.0	0.8

Error distributions obtained by id07 are investigated using the diagrams shown in Figure 9. Comparing the two histograms with those calculated on wind power, sharper and narrower distributions for both the Milano and Catania tests are seen, which implies a higher level of predictability of solar power in this exercise. Catania, in particular, shows a pretty symmetric distribution, with forecast errors being lower than 7.5% of *NP* 83% of the time. The distribution obtained for Milano is less symmetric and shows a higher number of negative errors. For Milano, errors lower than 7.5% are observed in 72% of the cases.

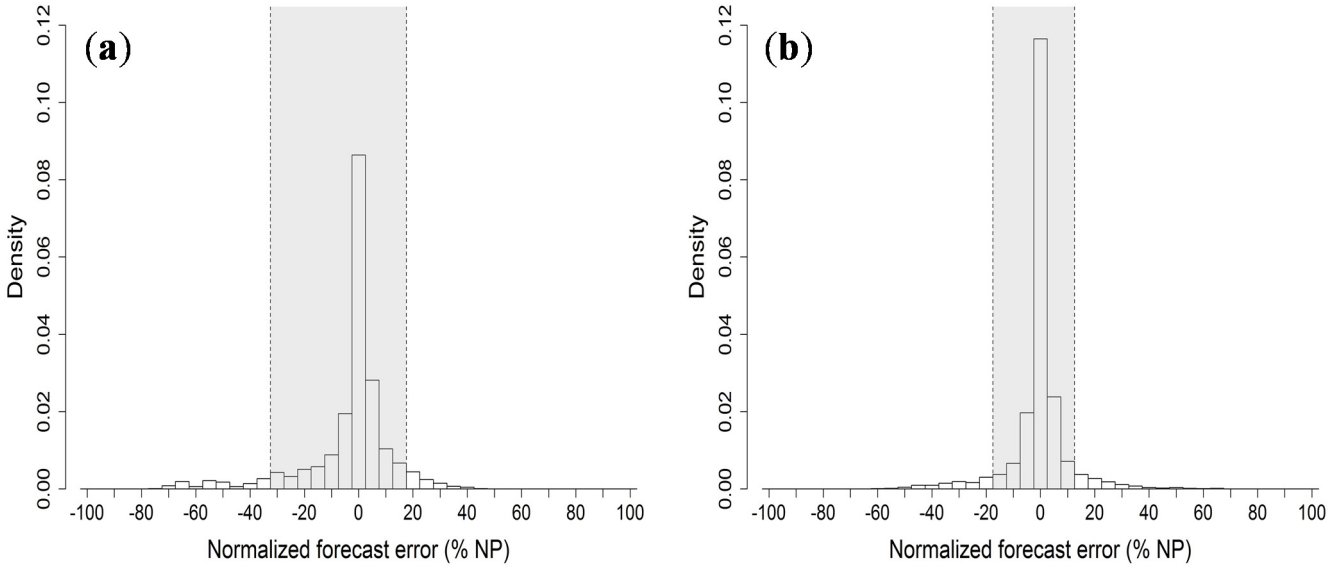

Figure 9. Error distributions (id07) of the best forecasts for the Milano case (**a**) and the Catania case (**b**). The grey area in each histogram delimits the 5%–95% quantiles interval.

Similarly to the wind part of the benchmark, one participant achieves the best result in both test cases in the solar application. For Milano, id07 used meteorological data provided by COST as input, then applying a quantile regression in order to estimate a clear sky production, a clear sky irradiance and a medium temperature [31]. A linear regression was also applied to explain the rate of observed clear sky production. The same method was used for Catania, with the additional step of performing a bias correction with a quantile regression, based on lead time and forecasted power. This last step was not applied to the Milano case due to the reduced amount of available data.

The other methods applied were mainly based on meteorological data provided by the organizers. A method proposed by id01 ranked third on Catania and second on Milano, using non-linear regression techniques such as random forests.

For Catania, the participant with the second best result (*i.e.*, id09) combined the output of the RAMS model provided by COST with the WRF ARW [32] model, initialized with boundary conditions from GFS. The forecasting system was (multiple) linear regression with several explanatory variables whose coefficients are estimated from data. Different variables derived from the NWP outputs were used in the regression, and they were fitted simultaneously in one consistent model. The derived variables were: direct component of the solar radiation on the tilted panel (30°, south) from WRF, diffuse component of the solar radiation on the tilted panel from WRF, difference between direct radiation on the tilted panel obtained from WRF and RAMS, difference between di use radiation on the tilted panel obtained from WRF and RAMS, interaction between direct WRF radiation and cosine of the zenith angle and interaction between diffuse WRF radiation and cosine of the zenith angle.

Other participants also applied linear regression techniques: id05 used all available values from GHI and solar power, applying linear regression to derive the relevant coefficients, but the results appear less effective; id08 forecasted the power output with a regression model based on the adjusted solar irradiance incident on the PV model surface and the solar cell temperature, which were calculated with the isotropic sky model and the standard formula with nominal operating cell temperature (NOCT) respectively, obtaining average results on both plants.

Few participants applied machine-learning techniques. ANN with a back-propagation algorithm was used by id06 with average results on Milano.

It should be noted that an SVM based application performed by id04 ranked third on Catania considering only the first prediction day. However, forecasts for the 24–48 h and 48–72 h ahead were missing.

4.4. Extension to Probabilistic Wind Power Forecasts

Probabilistic forecasts were provided by a limited number of participants only for the wind test cases, in terms of quantiles of the wind power PDF for each time-step. For Abruzzo, two participants provided 19 quantiles from 5% to 95% while the third one provided nine quantiles from 10% to 90%. As previously stated, the ranking is made using the *CRPS* index. Rank histograms are also presented to compare statistical consistency of the different ensemble forecasts. Finally, sharpness diagrams compare the relative forecast frequencies of the different forecasts for each test case.

Figure 10 shows the *CRPS* calculated at each of the 3-hourly, 0–72 h ahead time-steps for the Abruzzo case. As in the deterministic evaluation, the index is expressed as a percentage of both *NP* and *MP*. Looking at the diagram, the *CRPS* trends of all the participants look similar. A significant diurnal cycle is evident, showing better scores during morning hours. This reflects what is observed in the deterministic evaluation. The best result, expressed as average *CRPS* value over the entire 0–72 h forecast horizon, is achieved by id13 with 7.0% *CRPS*/*NP* and 39.4% *CRPS*/*MP*. Only for this trend line, bootstrap confidence bars are added in order to check for statistically significant differences between the models. It appears that, for most of the time-steps, the *CRPS* of id13 is not significantly better than the other participants, in particular during the worst performance periods during night hours.

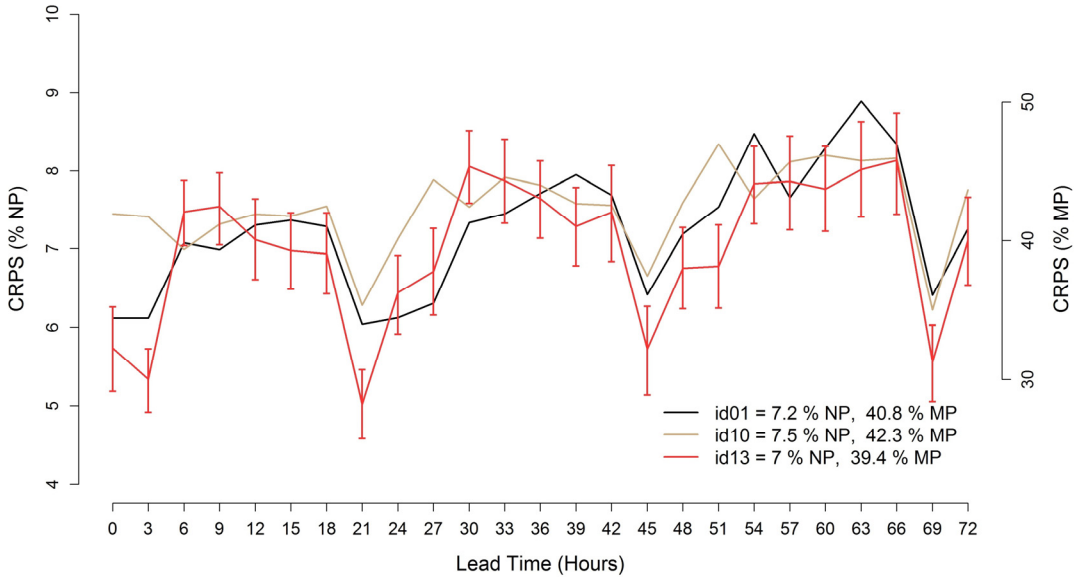

Figure 10. Error trends of the 0–72 h ahead probabilistic wind power forecasts starting at 12 UTC for the Abruzzo test case (*CRPS*/*NP* % and *CRPS*/*MP* %).

Statistical consistency is checked with the rank histograms reported in Figure 11.

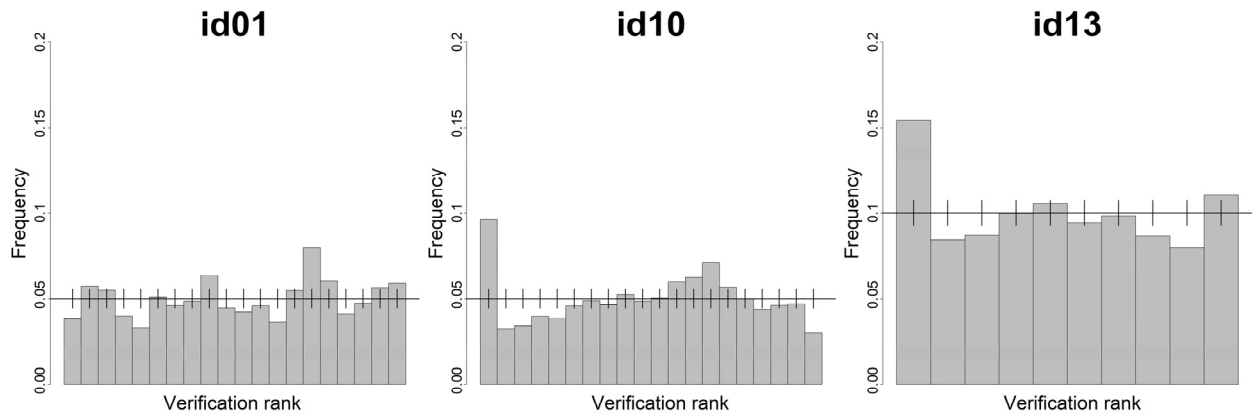

Figure 11. Rank histograms of the 0–72 h ahead probabilistic wind power forecasts starting at 12 UTC for the Abruzzo test case.

The different number of bins reflects the number of quantiles computed by the models: id13 delivered a wind power distribution made of 9 quantiles, while the other two participants used 19 quantiles. The vertical bars shown in the diagrams are calculated with a quantile function for a binomial distribution, in order to show a range in which deviations from the perfectly uniform distribution are still consistent with reliability. Deviations are in fact possible, since the number of samples in each bin is limited. The bars delimit the 5%–95% quantiles of the binomial distribution.

id01 shows a higher level of consistency. In fact id10 and id13 are slightly under-dispersive (*i.e.*, over-confident) having the first and last bins more populated. In each case, it can be seen that about half of the bins with deviations from the perfect frequency still lie in the consistency range delimited by the consistency bars.

Figure 12 shows the results for the Klim test case, for which 3-hourly, 0–48 h ahead power distributions were produced. In this case id10 performs better than id01. The boot strap confidence intervals show that the differences are statistically significant for every lead time. Figure 13 reports the rank histograms calculated for the two data sets.

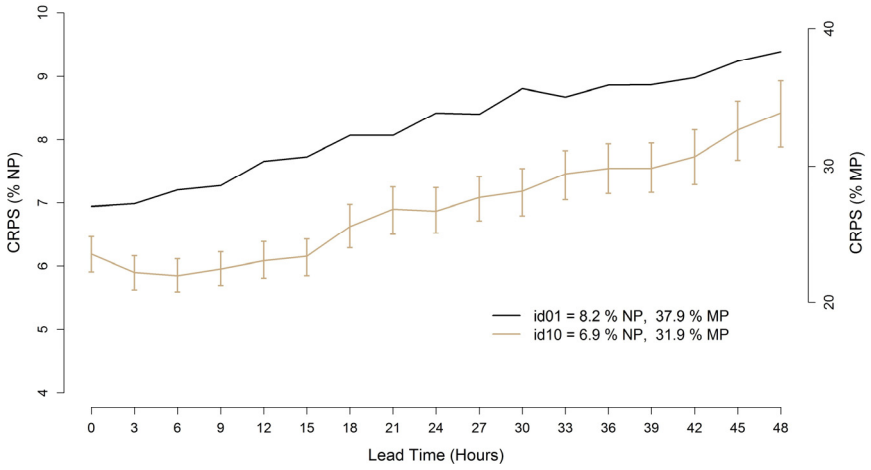

Figure 12. Error trends of the 0–72 h ahead probabilistic wind power forecasts starting at 0, 6, 12 and 18 UTC for the Klim test case (*CRPS*/*NP* % and *CRPS*/*MP* %).

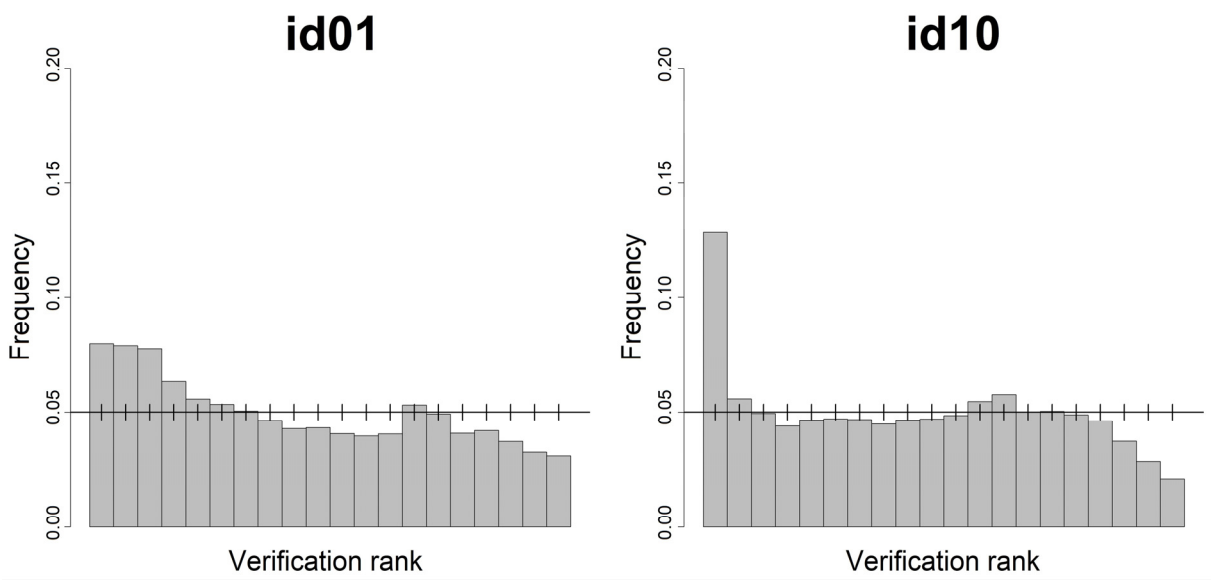

Figure 13. Rank histograms of the 0–72 h ahead probabilistic wind power forecasts starting at 0, 6, 12 and 18 UTC for the Klim test case.

Both models show a similar behavior and exhibit a positive bias. In fact, the first bins appear more populated than the others. This happens in particular for id10. However, the distribution obtained by id10 appears more consistent than that of id01 as demonstrated by the quantile bars, which maintain deviations from a uniform frequency within consistency for a greater number of bins.

Figure 14 shows sharpness diagrams for Abruzzo and Klim respectively. In the diagrams, the average of produced power is used as threshold value.

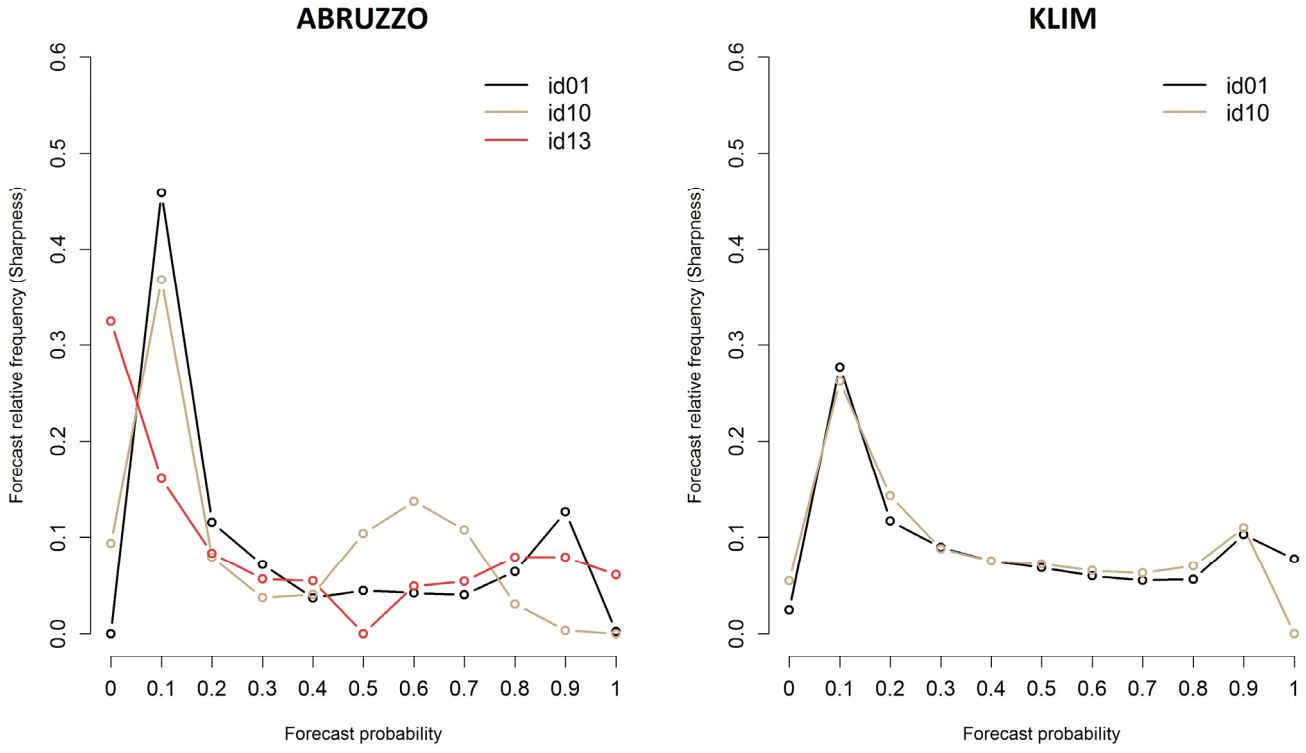

Figure 14. Sharpness diagrams for Abruzzo (**left**) and Klim (**right**). Mean produced power is used as threshold value.

In the case of Abruzzo, id13 shows sharper forecasts than the other participants. In fact, he is able to forecast both cases with probabilities equal to 0 and 1 with higher relative frequencies. In the case of Klim, the participants show a very similar trend. id01 behaves slightly better, in fact his relative forecast frequency for probability class equal to 1 is higher than that of id10.

The forecasting method used by id13 on Abruzzo is a local quantile regression with wind speed and wind direction as predictors [3]. On Klim, id10 applied conditional kernel density estimation with a quantile-copula estimator, using forecasted wind speed and direction, hour of the day and forecast lead time as inputs [33]. 5% to 95% quantiles were computed from the forecasted PDF using numerical integration.

5. Discussion

The benchmarking exercise gathered contributions from a wide range of countries, allowing raising some interesting discussion points. For the wind power prediction part, the best result reports a 9.0% *MAE*/*NP* (50.9% *MAE*/*MP*) on the Abruzzo test case, on a 0–72 h ahead forecast horizon. On the Klim case, a 9.5% *MAE*/*NP* (43.7% *MAE*/*MP*) is observed on a 0–48 h ahead forecast range.

As previously stated, the two power plants are located in very different locations, Abruzzo in a complex-terrain site in central Italy and Klim in a flat-terrain site in Northern Denmark. The predictability conditions are in principle more favorable for Klim. Furthermore, the representativeness errors of the meteorological model, mainly due to the more complex topography of the Abruzzo wind farm, mask the decreasing predictability related to the increasing forecast lead time, *i.e.*, the average *MAE* obtained by the participants on the first and on the third day ahead are quite similar.

However, it should be considered that the meteorological data provided for the Klim case were generated back in 2001 and 2002. As a consequence the *MAE*/*NP* obtained for Abruzzo, using more recent ECMWF global model data with 0.125° resolution for a complex-terrain site, followed by an effective post-processing system, is slightly lower than what was achieved for Klim. However, the Klim wind farm has a higher load factor than Abruzzo and subsequently the resulting *MAE*/*MP* value is around 7% lower for Klim.

The results for Klim are similar to those reported in [18]. A comparison between the current exercise and the results shown in [18] seems indicating that some improvements in the statistical post-processing techniques have been reached. Comparison should however be made with caution, since in [18] data availability for the wind farm was over 4 years, instead of 2 years as in the current exercise. The problem of having less recent and a lower spatial resolution in the forecast data for Klim was investigated by the organizers performing a higher resolution meteorological forecast. ECMWF reforecasts data with 0.25° horizontal resolution were retrieved for the 2-year period 2001–2002 and used as boundary conditions for a 4 km resolution run using RAMS. Then, a post-processing system based on an ANN was applied. The results show an improvement over the best result of the exercise, with a total *MAE*/*NP* about 0.5% lower. The same post-processing was also applied on the HIRLAM data for comparison, showing *MAE*/NP values similar to those obtained during the exercise. This allows concluding that in the Klim case, a source of improvement with respect to past forecasting systems is given by an increase in horizontal resolution of both the boundary conditions and the limited

area model. Another explanation is the better accuracy achieved by the ECMWF forecast system in the last 10 years, not only related to the spatial resolution, but also to other model developments.

For the solar power prediction part, the best result is a 7.0% *MAE*/*NP* (30.1% *MAE*/*MP*) on the Milano case, for a 3–72 h ahead forecast range. On the Catania case, a 5.4% *MAE*/*NP* (14.8% *MAE*/*MP*) is observed on a 1–72 h ahead forecast horizon.

Also in this case the power plants are located in different areas, with different predictability conditions. In the case of Milano there are higher levels of pollution in the suburban area surrounding the plant, in which the presence of aerosol particles influences solar radiation at the ground. Catania is located in Sicily, where solar irradiance availability is usually higher than in Northern Italy. In fact, the average GHI measured at the Milano power plant for the whole test period is equal to 165.6 W/m^2, against 203.7 W/m2 measured in the Catania plant. These facts imply better predictability conditions for Catania. The role of the Etna volcano is however relevant for Catania in terms of release of volcanic ash, which may influence power production and are difficult to consider in a forecasting system. Furthermore, concerning Milano, it should also be taken into account that the training period is just 6 months instead of a full year.

Solar energy appears to be more predictable than wind energy, as demonstrated by lower error values and narrower error distributions, especially for Catania. This is due to a higher predictability of solar radiation forcing in clear sky conditions. This higher predictability of solar PV farms compared to wind farms could not be generally confirmed for other climatic conditions, *i.e.*, northern Europe where the frequency of cloudy days could be higher than in Milano and Catania.

The probabilistic forecasting part of the exercise shows modest response, with just a few submissions for the wind test cases. In a way, this could be considered as an outcome of the exercise, indicating a still limited capability of the community to produce this kind of forecasts. Anyway, the results are coherent with those of the deterministic application in terms of performance quality dependence with the terrain complexity. In fact, a similar level of performance is observed between Abruzzo and Klim in terms of *CRPS* normalized by NP, while the best *CRPS*/*MP* obtained for Klim is about 7% lower than for Abruzzo due to the higher load factor of the Danish wind farm.

Finally, some considerations about computational times can be drawn. Typical time requirements to run a 3-day ahead forecast using a limited-area model are nowadays in the order of about 0.5 h or less, using parallel computing. The main source of computational expense in the modeling chain is certainly required to generate boundary conditions (e.g., using the ECMWF deterministic global model). Post-processing methods usually require a lower computational effort and can run on common CPUs, being able to deliver the final power forecast in just a few minutes. In case of probabilistic forecasting, additional computational effort is necessary when using ensemble meteorological models to generate multiple evolutions of a required atmospheric variable. Probabilistic forecasting based on statistical post-processing techniques applied on a single deterministic run (*i.e.*, quantile regression) requires similar computational time as that of a deterministic forecast. Some evidences of what mentioned are reported in [30].

6. Conclusions

In this paper, the benchmarking exercise organized in the frame of the COST Action "WIRE" is thoroughly described. The aim of this exercise was to establish the performance achievable by using state of the art models and tools for short-term wind and solar energy forecasting. Hence, it was possible to investigate the relationship between the weather dependent power production, which is by nature highly intermittent, and the energy distribution towards the end-users. In fact, the intermittence of these energy sources and the subsequent fluctuations, caused by their constantly increasing penetration into the electric grids, increase the importance of developing the best possible weather information and forecasting techniques to predict the energy production of wind and solar power plants.

Looking back to other comparison exercises, some conclusions about the evolution of performance can be drawn. The impression is that a substantial source of improvement in the last 10 years should be addressed to the NWP progress. In fact, testing the same post-processing technique on older, lower spatial resolution and on newer, higher spatial resolution meteorological data allowed getting better results. It is tricky to draw certain conclusions from a single test case, however an increase in model performance due to higher resolution and better accuracy appears reasonable. Also, improvements in the statistical modeling seem plausible, as noticeable by comparison with previous applications on the same model data.

Regarding the different modeling techniques, in the case of wind power the most effective methods proved to be based on machine learning algorithms, which can be applied effectively to deal with non-linear relations between weather parameters and power production. Fewer participants applied machine learning techniques in the solar forecasting contest, where linear regressions have been the most used. In particular, quantile regression provided the best performances despite a low computational effort.

In general, it is still difficult to find generalized answers from such an exercise. More work using more test cases, data and models needs to be performed in order to achieve a global overview of all possible situations. Test cases located all over Europe, the US and other relevant countries should be considered, trying to represent most of the possible meteorological conditions. Furthermore, standardized evaluation procedures shall be used when doing benchmarking studies, in order to facilitate comparisons between different test cases. Generally it could be quite difficult to find large amounts of data, especially dealing with confidentiality issues. Exchange of data and information should however be facilitated, in order to really reach a global view of the situation in the renewable energy forecasting field. Moreover, solar power data used for the exercise are of public domain and could be used by the community to test their models on the same dataset. Solar data may be obtained by contacting the main author.

Acknowledgments

This work has been partly financed by the Research Fund for the Italian Electrical System under the Contract Agreement between RSE S.p.A. and the Ministry of Economic Development-General Directorate for Nuclear Energy, Renewable Energy and Energy Efficiency in compliance with the Decree of 8 March 2006. The authors acknowledge Alain Heimo, chair of the COST Action

ES1002 “WIRE”, as well as René Cattin of Meteotest for their support. COST is acknowledged for sponsoring a Short Term Scientific Mission at DTU, in which part of the evaluation routines were developed. Dario Ronzio of RSE SpA is acknowledged for preparing solar power and meteorological data for Milano and Catania and wind power data for Abruzzo. The authors also acknowledge Sue Ellen Haupt of NCAR for useful suggestions to improve the manuscript.

Author Contributions

Simone Sperati developed evaluation routines, performed forecast evaluation and prepared the manuscript. Stefano Alessandrini performed additional forecast evaluation and contributed to the preparation of the manuscript. Pierre Pinson proposed evaluation routines and provided comments on the manuscript. George Kariniotakis provided data of previous benchmark exercises and commented on the manuscript. All of the authors read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Alessandrini, S.; Davò, F.; Sperati, S.; Benini, M.; Delle Monache, L. Comparison of the economic impact of different wind power forecast systems for producers. *Adv. Sci. Res.* **2014**, *11*, 49–53.
2. Molteni, F.; Buizza, R.; Palmer, T.N.; Petroliagis, T. The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. R. Meteorol. Soc.* **1996**, *529*, 73–119.
3. Bremnes, J.B. Probabilistic wind power forecasts using local quantile regression. *Wind Energy* **2004**, *7*, 47–54.
4. Brown, B.G.; Katz, R.W.; Murphy, A.M. Time series models to simulate and forecast wind speed and wind power. *J. Clim. Appl. Meteorol.* **1984**, *23*, 1184–1195.
5. Giebel, G.; Brownsword, R.; Kariniotakis, G.; Denhard, M.; Draxl, C. *The State-of-the-Art in Short-Term Prediction of Wind Power: A Literature Overview*, 2nd ed.; ANEMOS.plus: Risoe DTU, Roskilde, Denmark, 2011.
6. Monteiro, C.; Heko, H.; Bessa, R.; Miranda, V.; Botterud, A.; Wang, J.; Conzelmann, G. *Wind Power Forecasting: State-of-the-Art 2009*; Argonne National Laboratory: Lemont, IL, USA, 2009.
7. Cassola, F.; Burlando, M. Wind speed and wind energy forecast through Kalman filtering of Numerical Weather Prediction model output. *Appl. Energy* **2012**, *99*, 154–166.
8. Kou, P.; Gao, F.; Guan, X. Sparse online warped Gaussian process for wind power probabilistic forecasting. *Appl. Energy* **2013**, *108*, 410–428.
9. Wang, J.; Botterud, A.; Bessa, R.; Keko, H.; Carvalho, L.; Issicaba, D.; Sumaili, J.; Miranda, V. Wind power forecasting uncertainty and unit commitment. *Appl. Energy* **2011**, *88*, 4014–4023.
10. Jensenius, J.S.; Cotton, G.F. The Development and Testing of Automated Solar Energy Forecasts Based on the Model Output Statistics (MOS) Technique. In Proceedings of the 1st Workshop on Terrestrial Solar Resource Forecasting and on the Use on Satellites for Terrestrial Solar Resource Assessment, Washington, DC, USA, 2–5 February 1981.

11. Diagne, H.M.; David, M.; Lauret, P.; Boland, J. Solar irradiation forecasting: State-of-the-art and proposition for future developments for small-scale insular grids. In Proceedings of the World Renewable Energy Forum (WREF), Denver, CO, USA, 13–17 May 2012.
12. Heinemann, D.; Lorenz, E.; Girodo, M. Forecasting of solar radiation. In *Solar Energy Management for Electricity Generation from Local Level to Global Scale*; Dunlop, E., Wald, L., Suri, M., Eds.; Nova Publishers: New York, NY, USA, 2006; Chapter 7, pp. 83–94.
13. Pedro, H.T.C.; Coimbra, C.F.M. Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy* **2012**, *86*, 2017–2028.
14. Amrouche, B.; le Pivert, X. Artificial neural network based daily local forecasting for global solar radiation. *Appl. Energy* **2014**, *130*, 333–341.
15. Weron, R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *Int. J. For.* **2014**, *30*, 1030–1081.
16. Hong, T.; Pinson, P.; Fan, S. Global Energy Forecasting Competition 2012. *Int. J. For.* **2014**, *30*, 357–363.
17. Texier, O.; Girard, N. Wind power forecasting: A practical evaluation study of different wind power prediction services. In Proceedings of the European Wind Energy Conference and Exhibition, Marseille, France, 16–19 March 2009.
18. Kariniotakis, G.; Martí, I.; Casas, D.; Pinson, P.; Nielsen, T.S.; Madsen, H.; Giebel, G.; Usaola, J.; Sanchez, I.; Palomares, A.M.; *et al.* What performance can be expected by short-term wind power prediction models depending on site characteristics? In Proceedings of the European Wind Energy Conference (EWEC), London, UK, 22–25 November 2004.
19. COST Action ES 1002 WIRE: Weather Intelligence for Renewable Energies. Available online: http://www.wire1002.ch/ (accessed on 15 January 2013).
20. Benchmarking Exercise on Short-Term Forecasting Models for Renewable Generation. Available online: http://www.wire1002.ch/fileadmin/user_upload/Documents/ES1002_Benchmark_announcement_v6.pdf (accessed on 15 January 2013).
21. Madsen, H.; Pinson, P.; Kariniotakis, G.; Nielsen, H.A.; Nielsen, T.S. Standardizing the performance evaluation of short-term wind power prediction models. *Wind Eng.* **2005**, *29*, 475–489.
22. Pinson, P.; Nielsen, H.A.; Møller, J.K.; Madsen, H.; Kariniotakis, G. Nonparametric probabilistic forecasts of wind power: Required properties and evaluation. *Wind Energy* **2007**, *10*, 497–516.
23. Alessandrini, S.; Sperati, S.; Kariniotakis, G.; Pinson, P. WIRE COST Action, benchmark results. In Proceedings of the EWEA Technology Workshop: Wind Power Forecasting, Rotterdam, The Netherlands, 3–4 December 2013. Available online: http://www.ewea.org/events/workshops/wp-content/uploads/2013/12/EWEA-Forecasting-Workshop-2013-3-1-Stefano-Alessandrini-RSE.pdf (accessed on 20 January 2014).
24. Pielke, R.A.; Cotton, W.R.; Walko, R.L.; Tremback, C.J.; Lyons, W.A.; Grasso, L.D.; Nicholls, M.E.; Moran, M.D.; Wesley, D.A.; Lee, T.J.; *et al.* A Comprehensive Meteorological Modeling System—RAMS. *Meteorol. Atmos. Phys.* **1992**, *49*, 69–91.
25. Harrington, J.Y. The Effects of Radiative and Microphysical Processes on Simulated Warm and Transition-Season Arctic Stratus. Ph.D. Thesis, Colorado State University, Fort Collins, CO, USA, 1997.
26. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **2012**, *13*, 253–265.

27. Hersbach, H. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather For.* **2002**, *15*, 559–570.
28. Dobschinski, J. How good is my forecast? Comparability of wind power forecast errors. In Proceedings of the 13th International Workshop on Large Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Farms, Berlin, Germany, 11–13 November 2014.
29. Alessandrini, S.; Sperati, S.; Pinson, P. A comparison between the ECMWF and COSMO Ensemble Prediction Systems applied to short-term wind power forecasting on real data. *Appl. Energy* **2013**, *107*, 271–280.
30. Alessandrini, S.; Delle Monache, L.; Sperati, S.; Nissen, J.N. A novel application of an analog ensemble for short-term wind power forecasting. *Renew. Energy* **2015**, *76*, 768–781.
31. Bacher, P.; Madsen, H.; Nielsen, H.A. Online short-term solar power forecasting. *Sol. Energy* **2009**, *83*, 1772–1783.
32. Skamarock, W.C.; Klemp, J.B.; Dudhia, J.; Gill, D.O.; Barker, D.M.; Wang, W.; Powers, J.G. *A description of the Advanced Research WRF Version 2*; National Center for Atmospheric Research (NCAR): Boulder, CO, USA, 2005.
33. Bessa, R.J.; Miranda, V.; Botterud, A.; Zhou, Z.; Wang, J. Time-adaptive quantile-copula for wind power probabilistic forecasting. *Renew. Energy* **2012**, *40*, 29–39.